Reviews of submission #127: "DimScanner: A Relation-based Visual Exploration Approach Towards Data Dimension Inspection"

---------- Submission 127, Review 4 ----------

Reviewer: primary
Overall rating: 2.5 (scale is 1..5; 5 is best)

Topic Classification

Type: Technique and algorithm
Type: System
Topic: Visual representation and interaction technique

- Contribution to the field of Visual Analytics

The contribution is in the design of the tree structured view of views, which groups related views of dimensions of a high-dimensional dataset.

- The Review

The basic idea of visually grouping dimensions of a high-dimensional dataset to help analysts understand relationships, such as correlations, is good. The paper takes that idea a step further by not just grouping the dimensions, but grouping views (1D or 2D) of dimensions based on the dimensions contained in those views.

I think, though its not clear from the paper, that the advantage of this approach is that a single dimension can be represented in many views, and therefore can be grouped in several places in the visualization of views. Thus, the visualization presents a visual clustering of dimensions, where dimensions can be present in multiple clusters. Other than that, **its not really clear why you would want to group views of dimensions rather than group the dimensions directly.** In other words, what is the advantage of including the p-choose-2 2D scatterplots in the visualization instead of only the p histograms? This is especially unclear because all the examples in the paper show that the insights gained are about the dimensions, not about the views.

In fact, most of the examples are based on PCPlots of selected dimensions instead of using the histos and SPs. One possible answer is that the relations are computed based on the actual visuals, but I do not think this is the case, because the definition of MI in the paper seems to be based directly on the data dimensions, not the visual representations.

Unfortunately, the paper is very difficult to understand, and I am left with many open questions about how it works and why: What exactly are the users tasks that the system is designed for? What are they trying to learn about the dimensions? What are the kinds of relationships they can discover?

In 4.1, What specific MI implementation are you using? Is MI() == I()? Is this the same as the distance function suggested in 4.2? How does it compare to other metrics that have been used for

批注 [JS1]: Very important!
Design rationales: why 1D+2D instead of 1D
Response: In some cases views indicating combinational dimensions make more sense in reality. For example, combination of ``latitude'' and ``longitude'' is often more important than the individuals.

批注 [JS2]: Need to explain why it is so

批注 [JS3]: True

批注 [JS4]: Need to address them all in the paper

this same purpose (e.g. scagnostics, etc)?

Why are you using Quartets instead of some simpler method? Saying that it was used in Phylogenetics is not a reason. How does the Quartet method compare to other methods? You list several alternatives (MDS, SOM, Dendrogram), what is wrong with those?

In 5.0, what is novel here? I think the linking between the dimension SPLOM matrix and the tree view is a novel approach to exploring the space of SPs and Histos?

6 Are the colored highlights in Fig 4 automatic are user generated?

How are the specific green plots in Fig 4a selected out from among the many nearby (but dimmed) plots in the same branch in Fig 1? How does the user figure out that these 9 are the important ones in that branch?

Fig 6 "Because the dataset is dirty in terms of data contents, many one-dimensional views gather intensively on the same branch." Why?

Both examples are of relatively low dimensional data. How scalable is this visual representation?

In 7 user study, what is the "small multiples" layout that you are comparing to? Does it mean SPLOM? But, If so, SPLOMS are not scalable to high-dimensional data. Why are you comparing to that instead of something more relevant like alternative view metrics or view layout algorithms, such as Scagnostics/ScagExplorer?

What are the user tasks more specifically: "categorize the dimensions and identify their relations"? "three optional relation patterns are listed, including one- to-one correspondence, one-to-many correspondence and local correlation". What do these mean? If these tasks are important, they should be listing in the Introduction as the goal of your system design. Are there other types of relations?

Are there correct answers to these tasks? How are the answers derived? It looks like the correctness is defined on the same metrics that your visualization uses (IM)? If so, that seems a bit obvious that CTG should be better than SMG then, since it directly represents the information needed for the task, whereas SMG does not.

"ratio between the relation identification score to the classifications score" what does that mean?

8 "We demonstrate that a fair incorporation of quartet analysis effectively favors the cognition enhancement to the data dimensions." How? you would need compare it to an identical system minus the quartet analysis?

"Our system offers rich user controls, and enables data analysts to get the most context of an unfamiliar dataset." Most context?

批注 [JS5]: No, it is manually added to figure.

批注 [JS6]: They are 2d views of the dimensions in this branch

批注 [JS7]: explain more: the data are distributed non-uniformly
maybe not dirty…

批注 [JS8]: scalability discussion, in terms of number of dimensions as well as size of records

批注 [JS9]: The comparison is between tree-structured views and unstructured views
We are not able to compare/evaluate two techniques on the same platform.

批注 [JS10]: Explain them more.
Add them to the introduction section?

批注 [JS11]: It is. And it demonstrates the correctness of our c-tree structure.
Also, we encouraged our participants to list as many view-wise relations as possible. Thus, participants have different strategies in utilizing the c-tree structure.

批注 [JS12]: We did this in the user study, SMG.

批注 [JS13]: More context

If the main point of the paper is indeed the tree view visualization, then the Related Work should focus on the problem of dimension relationships. Expand 2.1, delete 2.2 and 2.3.

批注 [JS14]: Not sure about this.
The interactive controls are also one of our contributions.

The paper is written in a very dense language. After going through it several times, I think that the paper is actually not that complex, just the wording is much more complex than it needs to be.

批注 [JS15]: Re-wording

There may be some good ideas in the visualization of views, but this paper will require a major revision to make it acceptable.

Main Recommendation

Acceptable for presentation

The Summary Review

The reviewers have general consensus on the following issues:

Positives:

+ The use of category tree to visualize relations between dimensions, with some well-designed interactions, is a good idea.

+ The quartet tree is a novel approach to the problem.

Negatives:

- Needs a clearer description of the method, including metrics and algorithm.

- Needs more convincing arguments about its unique advantages in comparison to other closely related methods.

- User study does not reinforce the main contribution of category tree, since it compares to a very different control condition, and needs clarification of the study design and results.

- The writing needs to be cleaned up.

We recommend conditional acceptance based on the following required revisions:

- Eliminate the user study from the paper, using the saved space to make many clarifications in the rest of the paper as follows.

- Update the Related Work with more relevant references (several provided by reviews), focusing on dimension relations, dimension reduction.

- Clarify the method descriptions, including all the details of how it works, with examples. See the many questions in the reviews.

- Make a clear convincing argument about this method's unique advantages in comparison to other closely related methods for displaying dimension relations.

- Clean up the writing.

---------- Submission 127, Review 3 ----------

Reviewer: secondary
Overall rating: 3 (scale is 1..5; 5 is best)

Topic Classification

Type: System
Topic: Visual representation and interaction technique

Contribution to the field of Visual Analytics

The paper proposes a technique that presents a categorization tree for 1D and 2D views of multidimensional datasets, and an interactive system for exploring the relations between the dimensions. The scope of the paper is of interests to the visual analytics community. Studying efficient methods of analyzing multidimensional data via spatial layout is of high importance to visual analytics.

Expertise

3 (Knowledgeable)

The Review

The paper aims at addressing the relation discovery of multidimensional data. It proposes an interactive approach that presents a categorization tree (C-tree) of 1D and 2D views to the user, which allows the user to identify the relations between dimensions (e.g. one-to-many) quickly so as to find redundancy in dimensions, draw conclusions on data quality, etc. The paper implements the technique and its supportive interactions as a system called DimScanner. Case studies and user evaluations are presented to demonstrate the effectiveness of the proposed system and technique.

Overall, I think this is an interesting approach, especially in that the paper does not assume the user to have prior knowledge to the data in question, and the proposed system has been demonstrated to work with real data and users. The approach of applying quartet-based C-tree in view categorization is novel and justifiable. It looks promising in making analysis more accurate and less time-consuming compared with a simple set of coordinated multiple views. The supplemental video has shown that effective interactions help explore the C-tree. In this aspect I would argue for accepting this paper.

Nevertheless, a few concerns have given me a second thought. One problem is that the paper does not clearly describe its underlying C-tree construction technique and the relations between this work and previous work. The related work section could be refined to be more specific. The algorithm for constructing the C-tree is a bit confusing (see below). A second problem is that the paper does not compare the proposed technique against other ones that target at revealing the

dimensional relations with view layouts or user interactions. The advantages that potentially exist in a C-tree is not fully exploited, except for the node distances. Additionally, the discovered relations between the dimensions listed in the paper seem to be limited in basic relations as one-to-one, one-to-many, etc., which can be achieved by other methods as well.

The related work of this paper was a bit too general and could be improved. More particularly, the main contribution of the paper was not to address data quality and cleaning issues (though the mobile data case study appears to encounter this issue). Therefore section 2.1 is slightly off topic. The discussion of data exploration tools and coordinated views can be strengthened and made more detailed to describe the advantages of the proposed approach over the other visual analytics approaches for finding relations between dimensions. Additionally, a key aspect of this paper is on the usage of C-tree generated from quartet topology constraints. A brief survey of quartet and C-tree related techniques can be given, in addition to the references in Section 4. This will help the reader understand better why a C-tree is employed based on quartets here for seeking dimensional relations.

批注 [JS16]: Add a subsection of related work here

The algorithm described in Section 4 is of critical importance to the construction of the C-tree. However, the description of its construction process is lacking details and it is unclear to the reader how quartets are qualified. The paper describes that a quartet (ab|cd) should qualify if I(A;B) and I(C;D) are large enough and the other 4 relations are less significant. It seems to me that it would be sufficient to order the 6 relations and then check if the two largest relations are disjoint in views. Then we can use a threshold k to determine if the 3rd largest relation is sufficiently smaller than the 2nd largest relation. The description of Algorithm 1 is confusing. Line 1 ranks the relations and I assume this means sorting. However this would imply MaxIndex being 0 (or 5 if sorted increasingly). However line 2 still gets the index as if it would be not necessarily a fixed value. Line 3 indicates that the second largest relation would be at 5 - MaxIndex which does not make sense to me regardless of whether the relations are sorted or not. Maybe I am missing something here, but the description of the algorithm does not give me a clear understanding of the process. Intuitively I can understand that the main goal is to first identify a set of qualified quartets, avoiding the overwhelming size of O(m^4) candidates by cutting off at some threshold k. In addition, I(A;B) was computed from values in views A and B. The values of a view was not well defined in the paper. For a 2D view A, does D(A) include the values from both dimensions as a list/set? What happens if the view is 1D and how are a 1D view B matched with a 2D view A in terms of I(A;B)?

批注 [JS17]:
I will re-word this part

批注 [JS18]:

A more crucial missing part here is about justifying the unique advantages of a C-tree. The discovered relations, e.g. one-to-one and one-to-many, are possible to be revealed by other technique such as a scatterplot matrix (SPLOM), or simply a parallel coordinates plot (PCP). The discovery of such relations in this paper was via brushing and linking the visualized elements across multiple 1D/2D views. The user then verifies the relations in a PCP. What if the user directly uses a PCP and performs the verification step (sliding along an axis)? What if the user opens a SPLOM and performs hovering and linking interactions (single/multiple elements would be highlighted in the views in a same row/column)? I think the same relations could be found, assuming that the number of dimensions is not overwhelmingly large to prohibit the usage of PCP and SPLOM. IMO the proposed technique should have its advantages in reducing the number of dimensions to be

批注 [JS19]: Also mentioned in the primary review.

explored and avoiding the problem resulting from a large number of dimensions. This could have been implicitly shown in the case studies (100+ dimensions in the 911 dataset and mobile user profile dataset). However, I strongly encourage the paper to explicitly include a section that describes this advantage, which would make the paper much more convincing.

A comparison could be made against other works that directly address the relation discovery between dimensions. GPLOM, an extended variant of SPLOM, which also present multiple views for identifying relations between variables could be quite relevant.

J. F. Im, M. J. McGuffin and R. Leung, "GPLOM: The Generalized Plot Matrix for Visualizing Multidimensional Multivariate Data," in IEEE Transactions on Visualization and Computer Graphics, vol. 19, no. 12, pp. 2606-2614, Dec. 2013. doi: 10.1109/TVCG.2013.160

Some other works apply dimension-wise rendering and visualization. For example, the work by Yuan et al. listed below provides an interactive interface for the user to view MDS plots of both dimension and data subsets. This allows dimensions that are closely related to be visualized and interactively grouped together.

X. Yuan, D. Ren, Z. Wang, and C. Guo. “Dimension Projection Matrix/Tree: Interactive Subspace Visual Exploration and Analysis of High Dimensional Data”, in IEEE Transactions on Visualization and Computer Graphics (InfoVis'13), 19(12):2625-2633, 2013.

The technique proposed in this work does not involve direct visual encoding of dimensional relations except for the implicit spatial distance of a C-tree layout. Therefore the perspective slightly differs. Yet the paper reader may wonder how the proposed technique compares with such an approach that directly visualizes the relations between dimensions (instead of just plotting views on combinations of 1D/2D dimensions).

The paper could benefit from a thorough discussion about its design decision compared with those works.

To summarize, the paper does present interesting work that helps unravel the complicated relations between dimensions in multidimensional datasets using a spatial layout. Particularly, the application of quartet-based C-tree could be a novel approach that yields benefits in user perception. My feeling is mixed for this paper. Despite the good C-tree idea, I think the paper could benefit from a major revision to include a detailed and clearer description of the algorithm, more convincing arguments in C-tree’s unique advantages, and comparisons with other techniques with a close focus. Considering that, I would argue for a weak rejection.

A few minor issues:

The C-tree renders line chart thumbnails as if they are stacked graphs. Although the paper mentions that it is to make the lines prominent, I think this will introduce ambiguity and make

them not distinguishable from bar charts or histograms. I suggest using bolder lines and rendering line charts directly. In order to create view summary snapshots the user has to click one by one the views to be added (as shown in the video). This could be simplified by implementing a lasso selection, which would make summary creation much more convenient.

Some figures are unnecessarily large but not very informative. For example, Figure 3 can be shrinked to one column. This will leave more space for arguments and discussions of the proposed technique. Dataset list widget (Figure 1 (c)) is said to have collapsed after a dataset is chosen. Figure 1 does not properly show this widget and I think the paper could omit its description as it is not very important.

Typos:

Section 4.1 last line: a large MI >>in<< results in a large co-occurrence …

The first “in” should be removed.

Section 4.2 first line of last paragraph: We adopt the Quartet >>MacCut<<

(QMC) algorithm [28], …

---------- Submission 127, Review 1 ----------

Reviewer: external
Overall rating: 4 (scale is 1..5; 5 is best)

Topic Classification

Type: Technique and algorithm
Topic: Visual representation and interaction technique

Contribution to the field of Visual Analytics

This paper contributes a novel method for discovering and visualizing relationships between variables in high-dimensional data. The authors include a tool that adds data exploration capabilities to aid a user in comparing data dimensions, with an accompanying evaluation.

Expertise

4 (Expert)

The Review

The authors make a potentially strong contribution with this work. High-dimensional data analysis is a complex problem, and helping analysts sift through the many related dimensions and understand those relationships could be very helpful. The specific problem that forms the motivation, i.e. understanding the relationships between the dimensions, is an interesting formulation, but a citation about the fact that analysts encounter this specific problem would make the motivation much stronger. Also in the intro, the user study should be mentioned up front so that the reader expects that type of evaluation.

Though I believe in the work and the contribution, there are several opportunities to improve the writing. Here are some of my concerns about what could be communicated better:

I found the most clarifying example image to be Figure 7. At that point I finally could clearly see the utility of the trees. This could come much earlier. (It's also out of order layout-wise compared to the other figures.) Quartets are also discussed well before they are explained. In general, a pass through the paper to check that the first time something is referenced is when it is best explained would be helpful.

批注 [JS20]: Dendrogram might be a better description for C-tree. Introduce it earlier?

There are some things missing from the algorithm writeup, including a definition of k (it's in the paper but it comes a bunch later). The step labeled as getting the second best pair seems actually to be taking the pair that can sit opposite the first chosen pair in a quartet. Really, though, the algorithm for making the quartets does not need so much explanation. The algorithm that uses the quartets to build trees on the other hand should be explained a little so that the reader

understands the structures being built.

Because the focus is on exploring dimensions, related work should contrast your approach with traditional dimension reduction techniques like PCA and factor analysis. The wikipedia page on dimension reduction gives a good overview of the space.

批注 **[JS21]:** Method comparison

For the evaluation, it seems that the tree is being offered as a contribution, but it is not what's being tested in the evaluation. In general, I didn't understand what exactly is a category tree and what are its uses.

批注 **[JS22]:** Evaluation not well described?

Some more minor notes:

* Graphs can break transitivity if they're directed.
* There are a number of typos and mismatched verb conjugations to be fixed, e.g. in 4.1.
* What does "distribution in the view" mean?
* The 'stages' are not comparable because one is done by computer and the other is done by human, so it's strange to have them next to each other as if they're the same type of thing.
* Eqn 1 is not exactly clear - 'a' is a value in the data of view 'A', so what is the probability of a value? Does this assume each is guassian distributed and compute the likelihood of the particular value?
* I don't know if the second part of figure 2 helps much
* In the user evaluation - why is the comparison point small multiples?

---------- Submission 127, Review 2 ----------

Reviewer: external
Overall rating: 4 (scale is 1..5; 5 is best)
Topic Classification

Type: System
Topic: Visual representation and interaction technique

Contribution to the field of Visual Analytics

Overall, I think the manuscript clearly contributes to VAST. The idea of creating a graph based on MI between dimensions and their pair-wise correlations is clever. The simple tools recombined in a clever way makes this manuscript a valuable contribution.

The authors should include the missing related work. I would strongly recommend an English proof reading pass though. The user study needs to be reworked.

Nonetheless, I think these changes are manageable in the 'rebuttal' phase.

Expertise

3 (Knowledgeable)

The Review

DimScanner describes a VA system to explore relationships between dimensions in high-dimensional data analysis. It creates a categorization tree of 1D and 2D views and allows sophisticated interaction to navigate the visualization space.

The manuscript introduces the topic well. The related work seems to be missing an important paper [1] which allows searching the visualization space. In contrast, DimScanner is more an exploration tool.

The structuring stage (Sec.4) is well described. I have some minor improvements here:
- omit the 'Design Rational' subtitles
- elaborate more about why I(A;B) is close to +/- 1 for highly related views

Algorithm 1 clarity improvements:
- k is also an input variable
- indicate that relations is sorted by: relations = sorted([I(A;B), I(B;C),...])
- then: MaxRelation = relations[relations.length-1]
  SecondMaxRelation = relations[relations.length-2]

TestRelation = relations[relations.length-3]
qualified = (TestRelation < k * SecondMaxRelation)

The exploring stage (Sec 5) is clearly described. For ease of reading I would suggest to have small figures for the assistive widgets in the text columns so that referring back to figure 1 is not necessary. There is enough space when other figures are shrunk a bit (e.g. Fig 3). The selection of views on top of the matrix seems very intuitive. I couldn't find out if the matrix can be re-ordered?

批注 [JS23]: Cannot be re-ordered

The case studies (Seattle 911 and mobile user profile) are wise choices and communicate the power of the system well. I especially liked the example of the unbalanced tree which shows some quality control capabilities of the proposed method. But I was wondering why Fig.6 and Fig. 7 are traditional phylogenetic tree visualizations vs. Fig.4 where the tree is represented in a force directed layout. Can the user switch the tree layouts?

批注 [JS24]: Explain why two tree layout are introduced

批注 [JS25]: yes

Figures 4 and 5 would benefit from more caption text explaining the panels a-f.

The user study is probably the weakest part of the manuscript. It would clearly help to introduce the two competing techniques visually. I am not entirely sure what the small multiple layout is. The result section (7.3.) should be reworked - what is the score in the evaluation? What means 'better'?

Overall, I think the manuscript clearly contributes to VAST. The idea of creating a graph based on MI between dimensions and their pair-wise correlations is clever. The simple tools recombined in a clever way makes this manuscript a valuable contribution.

The authors should include the missing related work. I would strongly recommend an English proof reading pass though. The user study needs to be reworked.

Nonetheless, I think these changes are manageable in the 'rebuttal' phase.

[1]:
@article{2016-voyager,
 title = {Voyager: Exploratory Analysis via Faceted Browsing of
Visualization Recommendations},
 author = {Kanit Wongsuphasawat AND Dominik Moritz AND Anushka Anand AND
Jock Mackinlay AND Bill Howe AND Jeffrey Heer},
 journal = {IEEE Trans. Visualization \& Comp. Graphics (Proc. InfoVis)},
 year = {2016},
 url = {http://idl.cs.washington.edu/papers/voyager},
}